<머신러닝 기반의 창업 업종 추천 시스템>

멘토 박재범

멘티 홍익대학교 컴퓨터공학과 박준형, 박성춘, 이현수, 이유진

창업 시 필수적으로 필요한 시장조사를 편리하게 하기 위해 머신러닝 기법을 이용하여 상권을 분류하고 예측한다. 전체 모델은 Team 1(현수, 유진), Team 2(준형, 성춘) 두 팀으로 나누어 진행한다.

Team 2는 상권 데이터를 활용하여 K-means 클러스터링으로 클러스터를 생성하고 상권을 특징별로 grouping한다.

Team 1은 Team 2에서 결과로 나온 클러스터 Data를 Input으로 Random Forest를 사용하여 분류기(Model)를 만든다. 새로운 상권이 Input으로 들어왔을 때 어느 클러스터에 해당하는지 분류하는 기능을 한다. 이 모델은 클러스터 별 특징과 분류된 클러스터에 해당하는 업종을 통해 편리한 상권 분석을 할 수 있도록 도와준다.

1. 분석 요구 사항 정의서

* 상권 분석에 대한 요구사항
* 상권을 새로운 특성으로 분류한 후 Grouping이 필요하여 Clustering
* 분류한 cluster에 대해서는 명확한 근거 필요
* 새로운 상권이 어느 cluster에 해당하는지 분류 및 예측
* 새로 입점 시 해당 상권의 정보를 분석, 예측

1. 분석 데이터 정의서

2-1. 데이터 정의서

* 데이터의 출처
* 서울 열린 데이터 광장 – 서울시 우리마을가게 상권분석서비스
* 데이터 항목과 건수

|  |  |  |
| --- | --- | --- |
| 데이터 원본 이름 | 기간 및 건수 | 항목 |
| 서울시 우리마을가게 상권분석서비스  (상권-추정유동인구) | 2018년 9, 10월 | 총 유동인구 수,  연령대 별 유동인구 수 |
| 서울시 우리마을가게 상권분석서비스  (상권-상주인구) | 2016년 4월 – 2018년 10월 | 총 상주인구 수,  연령대 별 상주인구 수 |
| 서울시 우리마을가게 상권분석서비스  (상권-추정매출) | 2018년 9, 10월 | 평균 영업 개월 수, 점포 수 |
| 서울시 우리마을가게 상권분석서비스  (상권-점포) | 2018년 10월 | 점포 수, 유사 업종 점포 수,  개업율, 개업 점포 수, 폐업율,  폐업 점포 수 |
| 서울시 우리마을가게 상권분석서비스  (상권-지수지표) | 2018년 8, 9, 10월 | 상권 안정성, 활성도, 성장성,  지수/지표 구간 |

2-2. 최종 변수 정의서

2-1의 데이터 중 최종적으로 모델에 활용된 항목을 팀별로 입력 변수(X) 와 결과 변수(Y)로 나누어 정의한다. Team 2의 클러스터링 결과로 나온 클러스터(Y)를 Team 1에서 X변수로 활용하여 새로운 상권 데이터가 들어왔을 때 Team 2의 Y 값 중 어떤 클러스터에 해당하는 지를 결과값으로 예측한다.

* Team 2

|  |  |  |
| --- | --- | --- |
|  | 변수명 | 변수의 의미 |
| X변수  각 연령층의 매출비율과  유사업종수로  클러스터링 | X1: 10’s\_sales\_rate | 10대 연령층의 매출 비율 |
| X2: 20’s\_sales\_rate | 20대 연령층의 매출 비율 |
| X3: 30’s\_sales\_rate | 30대 연령층의 매출 비율 |
| X4: 40’s\_sales\_rate | 40대 연령층의 매출 비율 |
| X5: 50’s\_sales rate | 50대 연령층의 매출 비율 |
| X6: 60’s\_sales\_rate | 60대 연령층의 매출 비율 |
| X7: silmilar\_store\_number | 유사업종 수 |
| Y변수 | Cluster\_id | 각 군집에 해당하는 번호 |

* Team 1

|  |  |  |
| --- | --- | --- |
|  | 변수명 | 변수의 의미 |
| X변수  각 연령층의 매출비율과  유사업종수로  클러스터링 | X1: 10’s\_sales\_rate | 10대 연령층의 매출 비율 |
| X2: 20’s\_sales\_rate | 20대 연령층의 매출 비율 |
| X3: 30’s\_sales\_rate | 30대 연령층의 매출 비율 |
| X4: 40’s\_sales\_rate | 40대 연령층의 매출 비율 |
| X5: 50’s\_sales rate | 50대 연령층의 매출 비율 |
| X6: 60’s\_sales\_rate | 60대 연령층의 매출 비율 |
| X7: silmilar\_store\_number | 유사업종 수 |
| Y변수 | Cluster\_id | 각 군집에 해당하는 번호 |

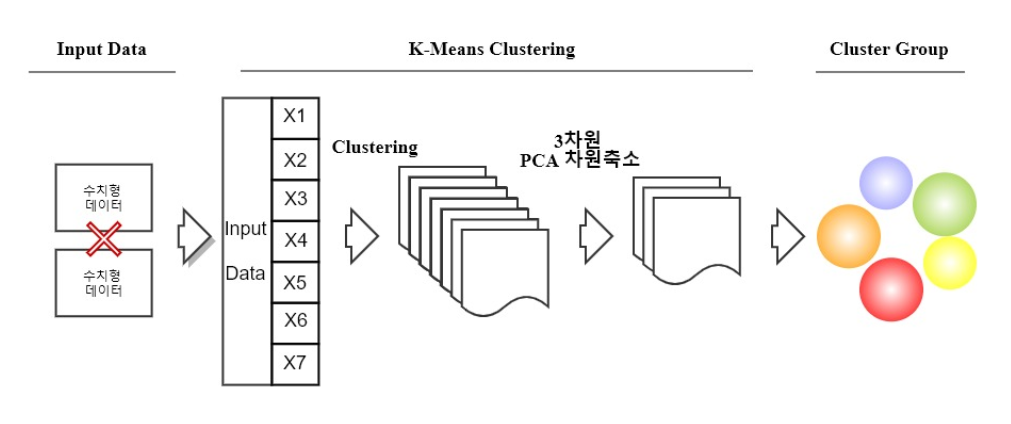
1. 분석 모형 정의서

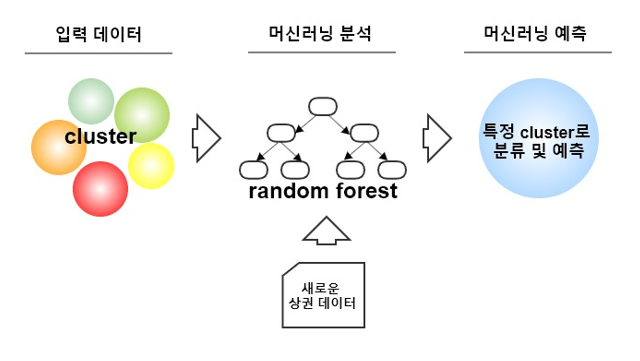
* 데이터 수집 -> 탐색적 분석 -> 변수 정의 -> 알고리즘 적용 ->

-> 학습 -> 비지도 학습: 해석, 지도 학습: 모형 정확도 해석 -> 시각화

3-1. 팀별 모형

* 클러스터링 모형(team 2)

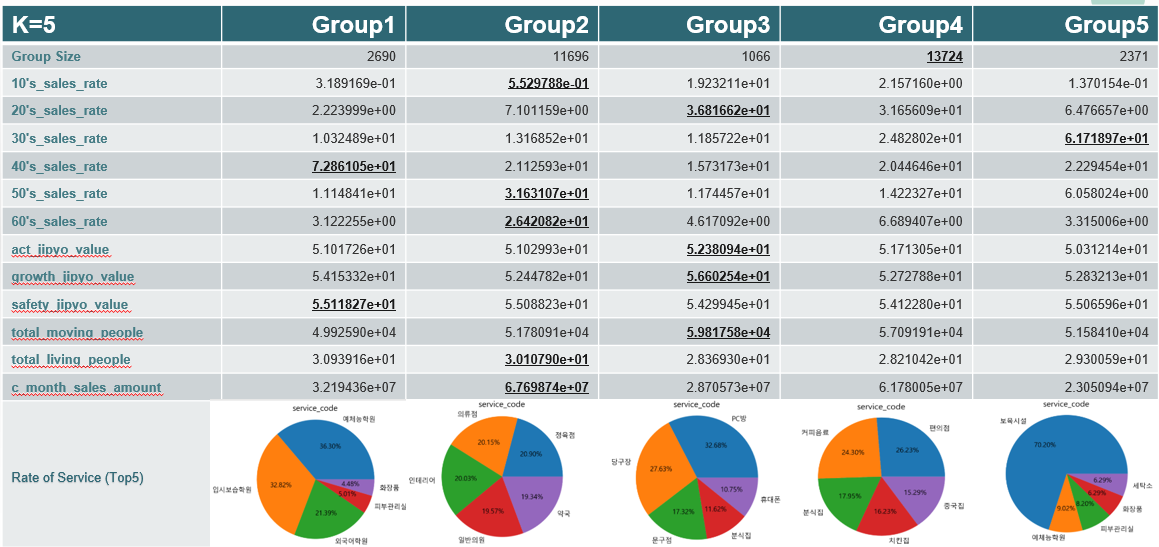


* 예측 모형(team 1)

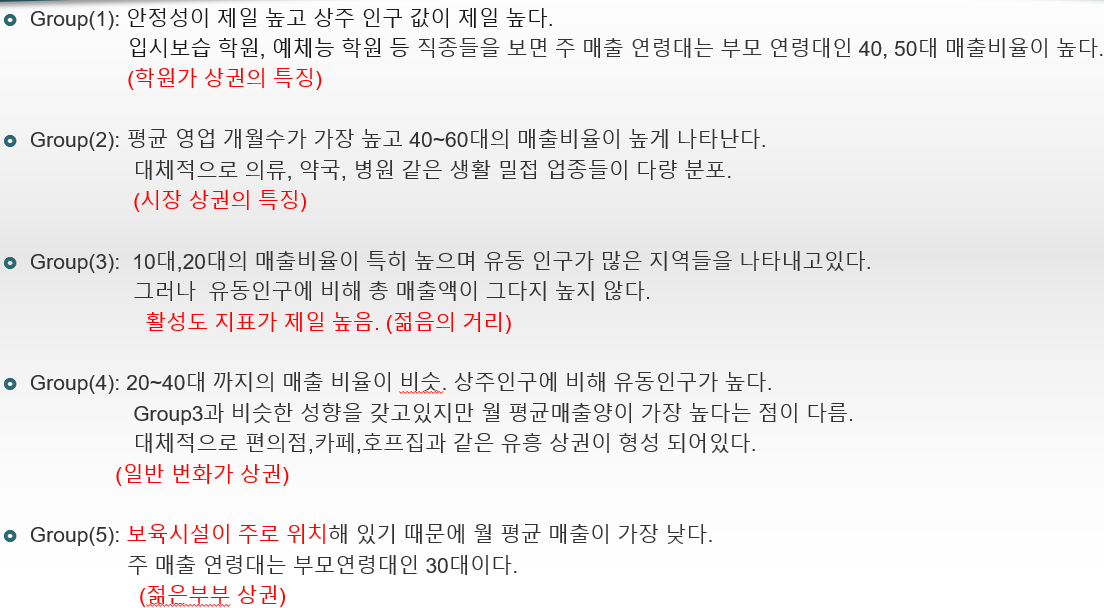
1. 분석 정의서

* 비지도 학습: Clustering (Team 2)

Scikit learn의 K-means clustering을 이용하여 클러스터링을 한 결과를 분석한 클러스터 별 의미는 아래와 같다.

* 1. 군집 분석(clusters Profiling)

**강조** - 가장 큰 값

* 1. 각 Group 특징
  2. Labeling
* Group(1): 학원가
* Group(2): 시장 상권
* Group(3): 젊음의 거리
* Group(4): 일반 번화가 상권
* Group(5): 젊은 부부 상권
  1. K-means Clustering 기법: K(군집의 수)값의 변화에 따른 군집 유형 분석

|  |  |
| --- | --- |
| K의 크기(군집의 개수) | K에 따른 군집 유형 분석 |
| K = 3 | 군집의 크기가 약 12000, 15000, 3000 개로  정확한 군집 별 특징을 알기 어려움 |
| K = 5 | 군집의 크기가 약 2500, 2000, 13000, 12000, 1000개로  군집 별 특징을 분석할 수 있음 |
| K = 7 | 몇몇 군집에 속한 데이터의 개수가 너무 적어  유의미하다고 보기 힘듬 |
| K = 3으로 두고  두 번 반복 | 약 12000, 15000 개로 나눠진 집단을  같은 조건(K=3)으로 clurstering 한 결과  군집 별 데이터의 개수가 더 편향적으로 나타남 |

결론: K=5 일 때 군집 별 크기의 편차가 작고 특징이 뚜렷해서 K=5 사용

* 지도 학습: 예측 정확도 (Team 1)

Random Forest를 이용하여 분류한 결과인 평가지표별 지표 값과 정오 분류표(Confusion Matrix)는 아래와 같다.

|  |  |  |  |
| --- | --- | --- | --- |
| 지표 유형 | 지표 | 지표설명 및 값 | |
| 정/오탐 유형 | TP(True Positive) | A를 A로 정 탐지 | |
| TN(True Negative) | B를 B로 정 탐지 | |
| FP(False Positive) | B를 A로 오 탐지 | |
| FN(False Negative) | A를 B로 오 탐지 | |
| 예측 정확도 | Accuracy  =(TP+TN)/(TP+TN+FP+FN) | 0.986 | |
| 재현율 | Recall = (TP)/(TP+FN) | Class 0 | 0.98 |
| Class 1 | 0.99 |
| Class 2 | 0.97 |
| Class 3 | 0.99 |
| Class 4 | 0.98 |
| 정밀도 | Precision = (TP)/(TP+FP) | Class 0 | 0.98 |
| Class 1 | 0.99 |
| Class 2 | 0.99 |
| Class 3 | 0.99 |
| Class 4 | 0.99 |

<Confusion Matrix>

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Confusion  Matrix | | 예측값 | | | | | |
| Class 0 | Class 1 | Class 2 | Class 3 | Class 4 | 합계 |
| 실제값 | Class 0 | 784 | 7 | 0 | 6 | 1 | 798 |
| Class 1 | 8 | 3412 | 4 | 38 | 1 | 3463 |
| Class 2 | 2 | 1 | 318 | 7 | 1 | 329 |
| Class 3 | 5 | 39 | 0 | 4105 | 4 | 4153 |
| Class 4 | 1 | 1 | 0 | 10 | 710 | 722 |
| 합계 | 800 | 3460 | 322 | 4166 | 717 | 9465 |

1. 전체 분석 시스템 구성도

(3.)에서 설명한 팀별 분석 모형을 구성도 형태로 표현하면 아래와 같다.

